Theoretical Computer Science 63 (1989) 1-17
North-Holland

AVOIDABLE PATTERNS ON TWO LETTERS*

Ursula SCHMIDT
Institut für Informatik, Universität Freiburg, D-7800 Freiburg, Fed. Rep. Germany

Communicated by D. Perrin
Received August 1987
Revised October 1987

Abstract. We examine avoidable patterns, unavoidable in the sense of Bean, Ehrenfeucht, McNulty (1979). We prove that each pattern on two letters of length at least 13 is avoidable on an alphabet with two letters. The proof is based essentially on two facts: First, each pattern containing an overlapping factor is avoidable by the infinite word of Thue-Morse; secondly, each pattern without overlapping factor is avoidable by the infinite word of Fibonacci. We further discuss the minimal alphabet on which very short patterns on a 2-letter and a 3-letter alphabet are avoidable.

1. Introduction

During the last years, several papers have considered combinatorial properties of words in connection with the occurrence of subwords of a special form. For example, some papers were devoted to the study of the generation of square-free or overlap-free words (Berstel [3, 5], Lothaire [7]), others to their determination and factorization (Crochemore [6], Main, Lorentz [8], Restivo, Salemi [10]). We are interested in a more general problem introduced in 1979 by Bean, Ehrenfeucht, McNulty [2], namely the study of so-called avoidable patterns.

It is possible to formulate the problem of unavoidable patterns as follows: We say that a word w divides a word u if there exists a nonerasing homomorphism h such that $h(w)$ is a subword of u. The word w is avoidable (respectively avoidable on an alphabet A with m letters) if there exists an infinite word on a finite alphabet (respectively on an alphabet A with m letters) which is not divisible by w. The opposite of avoidable is unavoidable (or in Zimin's terminology [15]: blocking). In this paper, we study the minimal alphabet A on which words on an alphabet E with two or three letters are avoidable.

If a pattern p on an n-letter alphabet is avoidable on an m-letter alphabet for some m, let $s(p)$ be the minimal such m. Bean, Ehrenfeucht, McNulty [2] and Zimin [15] gave implicit upper bounds for $s(p)$ which are exponential in n. The explicit bounds discussed in Baker, McNulty, Taylor [1] are linear in n. Nevertheless, this leads to $s(p) \leq 32$ for $n = 2$.

* This article was done while the author visited LITP, Université Pierre et Marie Curie, Paris.

We present a sharp bound for alphabets of cardinality 2: We show that in the case of patterns on a 2-letter alphabet E we have $s(p)=2$ for nearly all patterns. More precisely, we prove that each word on E of length at least $l=13$ is avoidable on a 2-letter alphabet. The proof is based essentially on two results: First, each pattern containing an overlapping factor is avoidable by the infinite word of Thue-Morse; secondly, each pattern without overlapping factors is avoidable by the infinite word of Fibonacci. Furthermore, we make some remarks about l, and we discuss $s(p)$ explicitly for short patterns p on a 2-letter and a 3-letter alphabet.

We present the main theorems in Section 3, which follows the definition section. Section 4 contains the proofs. In Section 5 we give some remarks on words on a 3-letter alphabet.

2. Preliminaries

We denote by A^* (respectively A^+) the free monoid (respectively semigroup) generated by the alphabet A. The elements of A^* and A^+ are called *words*. The cardinality of the alphabet A is denoted by $|A|$.

We say that w is a *subword* of the word u if there are words w_1 and w_2 such that $u=w_1ww_2$. We denote by $F(u)$ the set of all subwords of u. If $w=a_1a_2\ldots a_n$, with $a_i\in A$ for all $i=1,\ldots,n$, then $w^{\sim}=a_n\ldots a_1$ is the reversal of w. The number of occurrences of a letter a in a word w is defined by $|w|_a=|\{(w_1,w_2)\,|\,w=w_1aw_2,\ w_1,w_2\in A^*\}|$. The length of a word w is denoted by $|w|$. For any word w and any natural number k, w^k is defined so that w^0 is the empty word and $w^{k+1}=w^kw$.

Let A and E be alphabets. A homomorphism $h: E^*\to A^*$ is *nonerasing* if $|h(e)|\geq 1$ for each $e\in E$.

Let $p\in E^*$, $u\in A^*$. The word p *divides* u (or u is *divisible* by p) if there exists a nonerasing homomorphism $h: E^*\to A^*$ such that $h(p)$ is a subword of u. We also call p a *pattern* and $h(p)$ an *instance* of p.

Example ($E=\{x,y\}$, $A=\{a,b,c\}$).
- *ababaab* is an instance of *xxyx*, so *xxyx* divides *aababaabab*.
- x^2 does not divide *abcacbabcab*.

If p does not divide u, we say also that u *avoids* p. Clearly, if u avoids p, then each subword of u also avoids p.

If u avoids x^2 (respectively x^3), we say that u is *square-free* (respectively *cube-free*).

Let A be an alphabet with m letters. The pattern p is *avoidable on the m-letter alphabet* or *m-avoidable* if there exists an infinite set F of words on A which avoid p. We say that p is *avoidable* provided there exists a finite alphabet A such that p is avoidable on A. Otherwise, p is called *unavoidable*. In other words, p is unavoidable if each infinite set of words on a finite alphabet contains a word divisible by p.

An *infinite word* on A is a mapping $x: \mathbf{N} \to A$, denoted by $x = x_1 x_2 \ldots x_n \ldots$, where $x_n = x(n)$. So we can conclude that p is avoidable if and only if there exists an infinite word x on a finite alphabet such that x avoids p.

Example ($E = \{x, y\}$).

(1) xyx is unavoidable

(2) x^2 is not avoidable on $A = \{a, b\}$ since the only square-free words on A are a, b, ab, ba, aba and bab; but Thue [13] found an infinite square-free word on a 3-letter alphabet, so x^2 is avoidable.

(3) x^2 is 2-avoidable [13].

Clearly, if p is m-avoidable, then it is k-avoidable for all $k > m$.

Since the composition of two nonerasing homomorphisms is also a nonerasing homomorphism, the division is a transitive relation (it is also reflexive, but neither symmetrical nor antisymmetrical). This gives us the following property.

Property 2.1. *Let p and u be words on finite alphabets. If p divides u and p is m-avoidable, then u is also m-avoidable.*

In particular, all words containing a square are 3-avoidable. So we can state the following lemma.

Lemma 2.2. *Let $E = \{x, y\}$, $p \in E^*$. If $|p| \geq 4$, then p is avoidable.*

Bean, Ehrenfeucht, McNulty [2], and independently Zimin [15] showed that a doubled word is avoidable; a doubled word is defined as a word where each letter occurring in it occurs at least twice. One can affirm by induction on n that every word of length at least 2^n on an n-letter alphabet contains a doubled word of length at most 2^n as a subword. Therefore, we have the following general version of Lemma 2.2.

Theorem 2.3. (Bean, Ehrenfeucht, McNulty [2], Zimin [15]). *Let $p \in E^*$, $|E| = n$. If $|p| \geq 2^n$, then p is avoidable.*

Two other properties of avoidable words are easily verified.

Property 2.4. *A pattern $p \in E^*$ is avoidable if and only if $\sigma(p)$ is avoidable, where $\sigma: E \to E$ is an arbitrary permutation of the letters of E, or, more generally, $\sigma: E_1 \to E_2$ is an arbitrary bijection of the letters of E_1 onto the letters of E_2.*

Property 2.5. *A pattern p is avoidable if and only if its reversal $p^{\sim}$ is also avoidable.*

Therefore, we may consider avoidable or unavoidable patterns up to permutations and reversal.

We denote by $s: E^* \to \mathbf{N}$ the function defined by

$$s: E^* \to \mathbf{N}, \qquad s(p) = \min\{m \mid p \text{ is } m\text{-avoidable}\}.$$

For p unavoidable we set $s(p) = \infty$.

Example. $s(xyx) = \infty$, $s(x^2) = 3$.

Clearly, no word is 1-avoidable, so $s(p) \geq 2$ for all p and all E. Now, from Lemma 2.2 and the last two examples we get the following proposition.

Proposition 2.6. *Let $E = \{x, y\}$. If $p \in E^+$ is avoidable, then $s(p) \leq 3$.*

3. Results

Our aim is to investigate the patterns on the 2-letter alphabet $E = \{x, y\}$. In particular, we wish to know which of them are 2-avoidable, i.e., we are looking for patterns $p \in E^*$ with $s(p) = 2$.

We start this section with our main theorem which gives a very precise answer to this question. The rest of this section is composed of two parts: First, we give some results which allow to establish that "nearly all" words on E are 2-avoidable (Theorem 3.9); then we set out some other theorems such that the main theorem will be a consequence of them.

Let us state our main theorem.

Theorem 3.1. *All words $p \in E^+$, $E = \{x, y\}$, of length at least 13 are 2-avoidable.*

We pursue with some notations and results by Thue [13, 14]. Let $A = \{a, b\}$. The homomorphism $\mu: A^+ \to A^+$ is defined by

$$\mu(a) = ab, \qquad \mu(b) = ba.$$

When iterating this homomorphism, we get two sequences of words:

$$u_n = \mu^n(a), \qquad v_n = \mu^n(b).$$

These words are called *words of Thue–Morse.* They are related by the formulas

$$u_{n+1} = u_n v_n, \qquad v_{n+1} = v_n u_n \qquad (n \geq 0).$$

The first elements of the sequences are

$u_0 = a$, $v_0 = b$,

$u_1 = ab$, $v_1 = ba$,

$u_2 = abba$, $v_2 = baab$,

$u_3 = abbabaab$, $v_3 = baababba$,

$u_4 = abbabaabbaababba$, $v_4 = baababbaabbabaab$.

By further iteration, μ generates an infinite word $\mu^\omega(a)$ which is called the *infinite word of Thue-Morse* and is denoted by m:

$$m = \mu^\omega(a) = abbabaabbaababbabaababba\ldots.$$

Theorem 3.2 (Thue [13, 14]). *The word m avoids x^3 and $xyxyx$.*

Words avoiding the pattern $xyxyx$ are called *overlap-free*. So we can deduce the following corollary.

Corollary 3.3. *Each word containing an overlapping factor or a cube is 2-avoidable.*

Now it remains to examine overlap-free words. Restivo and Salemi [10] found a very precise factorization, for which we need some further notations.

Let $A_n = \{a_n, b_n\}$ be a pair of words on A^+ inductively defined by

$a_0 = a$, $b_0 = b$,

$a_{n+1} = a_n b_n b_n a_n$, $b_{n+1} = b_n a_n a_n b_n$ $(n \geq 0)$.

For example, we have

$a_1 = abba$, $b_1 = baab$,

$a_2 = abbabaabbaababba$, $b_2 = baababbaabbabaab$.

The elements of A_n are related to the words of Thue-Morse by the formulas

$$a_n = u_{2n}, \quad b_n = v_{2n} \quad (n \geq 0).$$

For $n \geq 0$, we define G_n (respectively D_n) as the set of left (respectively right) borders of order n:

$$G_n = \{1, a_n, b_n, a_n b_n, b_n a_n, a_n a_n, b_n b_n, a_n a_n b_n, a_n b_n a_n, b_n a_n b_n,$$
$$b_n b_n a_n, a_n a_n b_n a_n, a_n b_n a_n b_n, b_n a_n b_n a_n, b_n b_n a_n b_n\}$$

$$D_n = G_n^{\sim} = \{w^{\sim} \mid w \in G_n\}.$$

We note $A_n^i = \{w = w_1 \ldots w_i \mid w_k \in A_n,\ 1 \leq k \leq i\}$.

Now we can give the result of Restivo and Salemi [10].

Theorem 3.4. *Let $w \in E^*$ be an overlap-free word. There exists an integer $k \geq 0$ such that w can be factorized as follows:*

$$w = g_0 g_1 \dots g_{k-1} u d_{k-1} \dots d_1 d_0,$$

where $g_i \in G_i$, $d_i \in D_i$, $0 \leq i \leq k-1$ and $u \in \bigcup_{i=2}^{11} A_k^i$, and this factorization is unique.

This theorem states that the "central part" u of a finite overlap-free word is a subword of the infinite word m of Thue-Morse, whereas the "borders" have a different structure.

Since the words g_j, d_j, $0 \leq j \leq k-1$, and u have bounded length, we may state the following corollary, using the word u_i of Thue-Morse.

Corollary 3.5. *For all $i \in \mathbf{N}$, there exists a $c_i \in \mathbf{N}$ such that each overlap-free word on E of length at least c_i contains u_i as a subword.*

Recall that $u_{2i} = a_i$ and that $u_{2i+1} = a_i b_i$.

We define $\sigma_0 : A \to E$ to be the bijection

$$\sigma_0(a) = x, \qquad \sigma_0(b) = y.$$

Now, if we find i such that $\sigma_0(u_i)$ is 2-avoidable, then we can deduce that "nearly all" overlap-free words are 2-avoidable. Clearly, $i \geq 1$. As each word on A of length 11 contains an instance of $xyyx$, the pattern $\sigma_0(u_2)$ is not 2-avoidable, either.

We continue with $\sigma_0(u_3)$, and we are able to establish the following theorem.

Theorem 3.6. *The pattern $\sigma_0(u_3) = xyyxyxxy$ is 2-avoidable.*

In fact, this is a corollary to the following theorem.

Theorem 3.7. *The infinite word f of Fibonacci avoids $xyyxyxxy$.*

We use the following characterization of f: Let $A = \{a, b\}$. The homomorphism $\phi : A^+ \to A^+$ is defined by

$$\phi(a) = ab, \qquad \phi(b) = a.$$

The first elements of the sequence $\phi^n(a)$ are

$$\phi^0(a) = a,$$

$$\phi^1(a) = ab,$$

$$\phi^2(a) = aba,$$

$$\phi^3(a) = abaab,$$

$$\phi^4(a) = abaababa,$$

$$\phi^5(a) = abaababaabaab,$$

$$\phi^6(a) = abaababaabaababaababa.$$

By further iteration, ϕ generates an infinite word $\phi^\omega(a)$ which is called the *infinite word f of Fibonacci*:

$$f = \phi^\omega(a) = abaababaabaababaababaabaababaabaababaababa\ldots..$$

As a consequence of Corollary 3.5 and Theorem 3.6, we have this corollary.

Corollary 3.8. *Each overlap-free pattern $p \in E^+$ of length at least c_3 is 2-avoidable.*

We put Corollaries 3.3 and 3.8 together to give the following result.

Theorem 3.9. *For $|E| = 2$, each pattern on E^+ which is "long enough" is 2-avoidable.*

Now, in the second part of this section, we will specify what is "long enough", vaguely called c_3 in Corollary 3.8. Since $|g_i| = |d_i| \leq 4^{i+1}$ and $2 \cdot 4^k \leq 11 \cdot 4^k$, we could state without further reflections that

$$c_3 = \max\left\{|g_0| + |g_1| + |d_0| + |d_1| \,\middle|\, g_i \in G_i,\ d_i \in D_i,\ i = 1, 2, \text{ and } u \in \bigcup_{i=2}^{11} A_3^i\right\},$$

that is, $c_3 = 216$. This is a very large bound. So one should reason about the structure of the g_i, d_i and u for obtaining a smaller number.

We found it more interesting to consider the proof of Theorem 3.7 and to apply its method to another subword of the infinite word of Thue-Morse. This led to the following result.

Theorem 3.10. *The infinite word f of Fibonacci avoids xyyxxyyx.*

The word $\sigma^{-1}(xyyxxyyx) = abbaabba$ is a subword of the fourth word v_4 of Thue-Morse. We state as a consequence this corollary.

Corollary 3.11. *The pattern xyyxxyyx is 2-avoidable.*

Let $X = \{x^3, xyxyx, xyyxyxxy, (xyyx)^2\}$; $X \subset E^+$. The set X is a set of 2-avoidable words, and we know that each word on E which is "long enough" is divisible by an $x \in X$. So we may look for a minimal $l \in \mathbf{N}$ such that all words on E with length l are divisible by an $x \in X$. We are able to state the following theorem.

Theorem 3.12. *For all words $p \in E^+$ of length at least* 13 *there exists an $x \in X$ such that p is divisible by x.*

Now Theorem 3.1 is a consequence of Theorem 3.12.

Remark 3.13. It is not known whether this bound is optimal, i.e., whether there exists an $l \in \mathbf{N}$, $l < 13$, such that each $w \in E^+$ with $|w| \geq l$ is 2-avoidable.

The following statement will be proved in Section 5.

Proposition 3.14. $s(x^2yx^2) = 3$.

Corollary 3.15. *For the bound l in Remark* 3.13 *it holds that $l > 5$.*

4. Proofs

The interesting results are Theorems 3.7, 3.10 and 3.12. We start this section with recalling some properties of the subwords of the infinite word f of Fibonacci.

Let $E=\{x, y\}$, $A=\{a, b\}$; let $\phi: A^+ \to A^+$ be the homomorphism generating f, $F(f)$ the set of all subwords of f.

The set $\phi(A)=\{a, ab\}$ is a suffix code. So one can apply ϕ^{-1} to each word $u \in \{a, ab\}^*$. But $\phi^{-1}(u)$ is not necessarily a subword of f if u is a subword of f.

Example. $abaa \in F(f)$, $\phi^{-1}(abaa) = abb \notin F(f)$ and $ababa \in F(f)$, $\phi^{-1}(ababa) = aab \in F(f)$.

We can state the following proposition.

Proposition 4.1. *Let $u \in F(f)$, $|u| \geq 2$. If $u = au'b$, then $\phi^{-1}(u) \in F(f)$.*

Proposition 4.2 (Séebold [12]). *The words b^2, a^3, babab, aabaabaa are not subwords of f.*

A word $w \in A^*$ is a *conjugate* of a word $w' \in A^*$ if there exist $u, v \in A^*$ such that $w = uv$ and $w' = vu$.

Proposition 4.3 (Séebold [12]). *If $u^2 \in F(f)$, then there exists a $u' \in A^+$ such that u is a conjugate of u', and there exists an $n \in \mathbf{N}$ such that $u' = \phi^n(a)$.*

A complete proof of this proposition is given in [11].

Proposition 4.4. (Berstel [4]). *If $u \in F(f)$, then $u^{\sim} \in F(f)$.*

Now we will give the proof of Theorem 3.7, making use of Propositions 4.1 to 4.4.

Proof of Theorem 3.7. We will show that the following claim holds.

Claim. *The infinite word f does not contain a word of the form $w = uvvuvuuv_1$, where v_1 is the prefix of v of length $|v|-1$, i.e., v_1 equals the word v without its last letter.*

This claim implies that f does not contain a word of the form $uvvuvuuv$; that means f avoids $xyyxyxxy$.

Proof of the claim. Suppose that f contains a word of the form $uvvuvuuv_1$. Hence, u^2, v^2 and $(uv)^2$ are supposed to be subwords of f. Proposition 4.3 implies that there exist $r, s, t \in \mathbf{N}$ such that

$$|u| = |\phi^r(a)|, \qquad |v| = |\phi^s(a)|, \qquad |vu| = |\phi^t(a)|.$$

Hence $t = r+s$, and $s = r+1$ or $s = r-1$.

Without loss of generality, let $|u| = |\phi^r(a)|$, $|v| = |\phi^{r-1}(a)|$. Now, the proof is made by induction on r.

If $r=2$, i.e., $|v|=1$ and $|u|=2$, then $v=a$ by Proposition 4.2. Let $u=a_1a_2$, $a_1,a_2\in A$; then

$$w=a_1a_2aaa_1a_2aa_1a_2a_1a_2.$$

Since $a^3\notin F(f)$, we get $a_1=a_2=b$. One deduces that $b^4\in F(f)$; this is a contradiction to Proposition 4.2.

This establishes the basis of induction. Now we explicitly state the induction hypothesis.

Hypothesis of induction. *For all* $k\in\mathbf{N}$, $k\leqslant r$, *and for each pair of words* (u, v), *where* $|uv|=|\phi^{k+1}(a)|$ *and* $|u|=|\phi^k(a)|$, *we have that* $w=uvvuvuuv_1$ *is not a subword of* f, *where* v_1 *is the prefix of* v *of length* $|v|-1$.

Now we prove the claim for $r+1$ by contradiction. Suppose $w=uvvuvuuv_1\in F(f)$, where $|u|=|\phi^{r+1}(a)|$, $|v|=|\phi^r(a)|$. We distinguish four cases depending on whether the first letter of u respectively v is a or b.

In what follows, we distinguish, in the first case, again four subcases depending on whether the last letter of u respectively v is a or b; next, we reduce Case 2 to Case 1 by taking the reversal; and, finally, we bring Cases 3 and 4 back to Case 1 by shifting w with one letter to the left.

Case 1: *u and v begin with letter a.*

(α) *u and v end with letter a.* Let $u=au'a$, $v=av'a$; then

$$w=au'aav'aav'aa\underbrace{u'aav'aa}_{1}u'aau'aav'.$$

The subword indicated by 1 implies that u' and v' end with b. Set $\bar{u}=aau'$, $\hat{u}=au'$, $\bar{v}=aav'$. Now

$$w=\hat{u}\bar{v}\bar{u}\bar{v}\bar{u}\bar{u}\bar{v}. \qquad (*)$$

Hence, w begins with a and ends with b, as $\hat{u}$, $\bar{v}$, $\bar{u}$ do. Hence, these words satisfy the conditions of Proposition 4.1; it follows that

$$\phi^{-1}(w),\ \phi^{-1}(\hat{u}),\ \phi^{-1}(\bar{v}),\ \phi^{-1}(\bar{u})\in F(f).$$

Set $z_1=\phi^{-1}(\hat{u})$, $z=\phi^{-1}(\bar{u})$, $z'=\phi^{-1}(\bar{v})$. One has

$$\phi^{-1}(w)=z_1z'z'zz'zzz',$$

where z_1 is the suffix of z of length $|z|-1$. According to Proposition 4.4, $(\phi^{-1}(w))^{\sim}\in F(f)$; furthermore, Proposition 4.3 implies $|z|=|\phi^r(a)|$ and $|z'|=|\phi^{r-1}(a)|$; this is a contradiction to the induction hypothesis.

(β) *u ends with a, v with b.* Hence, $u=au'a$, $v=av'b$, $v_1=av'$, and v' ends with a. We have $|v'|>0$; otherwise, *babab* would be a subword of w, hence of f, a contradiction to Proposition 4.2.

(i) If $wb\in F(f)$, we have

$$wb=au'\underline{aav'}bvuva\underline{u'aau'}av'b.$$

So let $u' = bu''b$, $v' = bv''$; hence,

$$wb = \underline{abu''b}a(\underline{abv''b})^2\underline{abu''b}\underline{aabv''b}(\underline{abu''b}a)^2\underline{abv''b}.$$

It follows that $\phi^{-1}(wb) \in F(f)$. Set $z = \phi^{-1}(abu''ba)$, $z' = \phi^{-1}(abv''b)$, then

$$\phi^{-1}(wb) = zz'z'zz'zzz'.$$

Proposition 4.3 implies that $|z| = |\phi^r(a)|$ and $|z'| = |\phi^{r-1}(a)|$. But it follows from the induction hypothesis that $zz'z'zz'zzz' \notin F(f)$, hence $\phi^{-1}(wb) \notin F(f)$, a contradiction.

(ii) If $wa \in F(f)$, let $u' = bu''b$; hence, $u = abu''ba$. Let $v' = v''a$; hence, $v = av''ab$ and $v_1 = av''a$. We have $|v| \geqslant 3$; hence, $|v''| \geqslant 0$. It follows that

$$wa = uva\underline{v''ababu''b}avuua\underline{v''aa}.$$

If $|v''| = |u''| = 0$, then a^3 is a subword of f as a suffix of wa; this is a contradiction to Proposition 4.2. If $|v''| > 0$, and hence the last letter of v'' is b, one has $babab$ as a subword of f, a contradiction.

(γ) *u ends with b, v ends with a.* Hence, $u = au'b$, $v = av'a$, and

$$w = au'ba\underbrace{v'aa}_{2}v'aau'bav'\underbrace{aau'b}_{1}au'bav'$$

1 implies $u' = bu''a$, hence $u = abu''ab$;

2 implies $v' = v''b$, hence $v = av''ba$.

So, $\phi^{-1}(w) \in F(f)$. Set $z = \phi^{-1}(u)$, $z_1' = \phi^{-1}(av')$, $z' = \phi^{-1}(v)$. The word z_1' equals z' without its last letter. Now,

$$\phi^{-1}(w) = zz'z'zz'zzz_1',$$

where $|z| = |\phi^r(a)|$, $|z'| = |\phi^{r-1}(a)|$; this contradicts the induction hypothesis.

(δ) *u and v end with b.* Hence, $u = au'b$, $v = av'b$. Now, $u = au''ab$, $v = av''av$, and $v_1 = av''a$.

(i) If $wb \in F(f)$, then $\phi^{-1}(wb) \in F(f)$, and $\phi^{-1}(wb)$ has the form $zz'z'zz'zzz'$, where $|z| = |\phi^r(a)|$ and $|z'| = |\phi^{r-1}(a)|$, a contradiction.

(ii) If $wa \in F(f)$, then v'' ends with b since $v''aa$ is a suffix of wa. Set $z = \phi^{-1}(u)$, $z' = \phi^{-1}(v)$, and $z_1' = \phi^{-1}(av'')$. Then z_1' equals z' without its last letter. We have

$$w' = uvvuvuuav'',$$

and $\phi^{-1}(w') \in F(f)$. Furthermore,

$$\phi^{-1}(w') = zz'z'zz'zzz_1'$$

and $|z| = |\phi^r(a)|$ and $|z'| = |\phi^{r-1}(a)|$, a contradiction to the induction hypothesis.

Case 2: u begins with letter a, v with letter b. Now v^2, $uv \in F(f)$, hence u, v end with a. Let $u = au'a$, $v = bv'a$; then

$$w = au'abv'abv'aau'ab\underline{v'aau'aa}u'abv'.$$

It follows that u' and v' end with letter b. Set $\hat{u} = au$, $\bar{u} = aau'$, $\bar{v} = abv'$; then

$$w = \hat{u}\bar{v}\bar{u}\bar{v}\bar{u}\bar{u}\bar{v},$$

where $\hat{u}$, $\bar{u}$ and $\bar{v}$ begin with letter a and end with letter b. These are the conditions of case (*) (see Case 1(α)).

Case 3: *u begins with letter b, v with letter a.* Now vu, u^2 are subwords of w; hence, vu, $u^2 \in F(f)$; it follows that u and v end with a. Let $u = bu'a$ and $v = av'a$; then

$$w = bu'aav'aav'abu'aav'abu'abu'aav'.$$

Hence, $aw \in F(f)$. Set $\bar{u} = abu'$, $\bar{v} = aav'$; then

$$aw = \bar{u}\bar{v}\bar{u}\bar{v}\bar{u}\bar{u}\bar{v},$$

where $\bar{u}$, $\bar{v}$ begin both with letter a. So this is Case 1.

Case 4: *u and v begin with letter b.* It follows that u and v end with letter a, and $aw \in F(f)$. Let $u = bu'a$, $v = bv'a$; now, set $\bar{u} = abu'$, $\bar{v} = abv'$; then

$$aw = \bar{u}\bar{v}\bar{u}\bar{v}\bar{u}\bar{u}\bar{v},$$

where $\bar{u}$ and $\bar{v}$ begin with letter a. So this is also Case 1. □

Proof of Theorem 3.10. We will show that the following claim holds.

Claim. *The infinite word f does not contain a word of the form* $w = uvvuuvvu_1$ *where* u_1 *is the prefix of u of length* $|u|-1$.

This claim implies that f does not contain a word of the form $(uvvu)^2$; that means f avoids $(xyyx)^2$.

Proof of the claim. Suppose that f contains a word of the form $uvvuuvvu_1$. In particular, u^2 and v^2 are supposed to be subwords of f. It follows by Proposition 4.3 that there exist $r, s \in \mathbf{N}$ such that

$$|u| = |\phi^r(a)| \quad \text{and} \quad |v| = |\phi^s(a)|.$$

This time, there exists no relation between $|u|$ and $|v|$ with respect to $|\phi^r(a)|$, so the proof is made by induction on r and s.

(i) If $r = 1$ and $|v| = |\phi^s(a)|$, then $u = a$ and $w = avvaavv$. Proposition 4.2 implies that v begins and ends with b. Since v^2 is a subword of w, so $b^2 \in F(f)$; a contradiction to Proposition 4.2.

(ii) By symmetry, the case $s = 1$ and $|u| = |\phi^r(a)|$ leads also to a contradiction.

This is the basis of the induction.

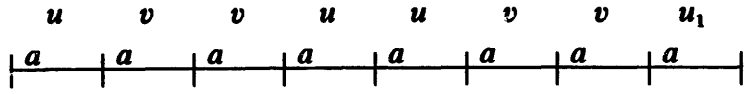
Hypothesis of induction. *For all integer* $k_1, k_2 \leqslant r$, *s and for all pairs of words* (u, v), *where* $|u| = |\phi^{k_1}(a)|$ *and* $|v| = |\phi^{k_2}(a)|$, *the word* $w = uvvuuvvu_1$ *is not a subword of f, where* u_1 *is the prefix of u of length* $|u|-1$.

Now we prove the claim for $r+1$ and s by contradiction.

Suppose that $w = uvvuuvvu_1 \in F(f)$, where $|u| = |\phi^{r+1}(a)|$, $|v| = |\phi^s(a)|$.

Case 1: *The words u and v begin with letter a.* We may shorten this part of the proof remembering the reasoning of the proof of Theorem 3.7.

One factorizes w before letter a to form $\phi^{-1}(w)$:



Hence, $\phi^{-1}(w) = zz'z'zzz'z'\hat{z}$, where $z = \phi^{-1}(u)$, $z' = \phi^{-1}(v)$ and $\hat{z} = \phi^{-1}(u_1)$. Since $|u| = |\phi^{r+1}(a)|$ and $|v| = |\phi^{s}(a)|$, one has $|z| = |\phi^{r}(a)|$ and $|z'| = |\phi^{s-1}(a)|$. It remains to show that the prefix z_1 of $\hat{z}$ of length $|z|-1$ is a prefix of z and that $\phi^{-1}(z_1) \in F(f)$.

- The word u does not end with a^2 since $a^3 \notin F(f)$.
- If u ends with ba, then z ends with ab, and u_1 ends with b. It follows that $\hat{z} = \phi^{-1}(u_1) \in F(f)$, and $\hat{z}$ ends with a. Hence, $\hat{z}$ equals z without its last letter, and $z_1 = \hat{z}$.
- If u ends with ab, then z ends with a and u_1 ends with a. Now it is sufficient to consider the lengths of the words. We have $|z| = |\hat{z}|$, and the prefix of z of length $|z|-1$ equals the prefix z_1 of $\hat{z}$ of length $|z|-1$. Let $u = u'ab$. Now $z_1 = \phi^{-1}(u') \in F(f)$.

Case 2: *The word u begins with a, v with b.* Since v^2, $uv \in F(f)$, the words u and v end with b. *Let* $u = au'a$, $v = bv'a$; then

$$w = au'abv'abv'aau'aau'abv'abv'aau'.$$

Hence, u' and v' end with b. Set $z_1 = \phi^{-1}(au')$ and $z' = \phi^{-1}(abv')$. Then

$$\phi^{-1}(w) = z_1z'z'bz_1bz_1z'z'bz_1 = z_1z'z'zzz'z'z,$$

where $z = bz'$, and z' equals z without its first letter. If $\phi^{-1}(w) \in F(f)$, then $(\phi^{-1}(w))^{\sim} \in F(f)$ by Proposition 4.4. This contradicts the hypothesis.

Case 3: *The word u begins with b, v with a.* Hence, $aw \in F(f)$, and u and v end with a. Let $u = bu'a$, $v = av'a$. Set $\bar{u} = abu'$, $\bar{v} = aav'$. Then

$$aw = (\bar{u}\bar{v}\bar{v}\bar{u})^2,$$

where $\bar{u}$ and $\bar{v}$ both begin with letter a. But this is Case 1.

Case 4: *u and v begin with the letter b.* Then $aw \in F(f)$. Set $\bar{u} = abu'$ and $\bar{v} = abv'$; then continue as in Case 3.

In an analogous way, one shows that the claim is true for $|u| = |\phi^{r}(a)|$, $|v| = |\phi^{s+1}(a)|$. □

Proof of Theorem 3.12. We want to show that all words $p' \in E^+$ of length 13 are divisible by some $p \in X$.

The set $X = \{x^3, xyxyx, xyyxyxxy, (xyyx)^2\} \subset E^+$ is a set of 2-avoidable words; furthermore, we know that each word which is "long enough" is divisible by some $p \in X$. Now we will look for the maximal length l of words on E not divisible by an element of X.

Suppose we have determined all patterns on E of length k avoiding all elements of X. According to Property 2.1, it is sufficient to extend these words in order to obtain the words of length $k+1$ avoiding each element of X. Besides, since we

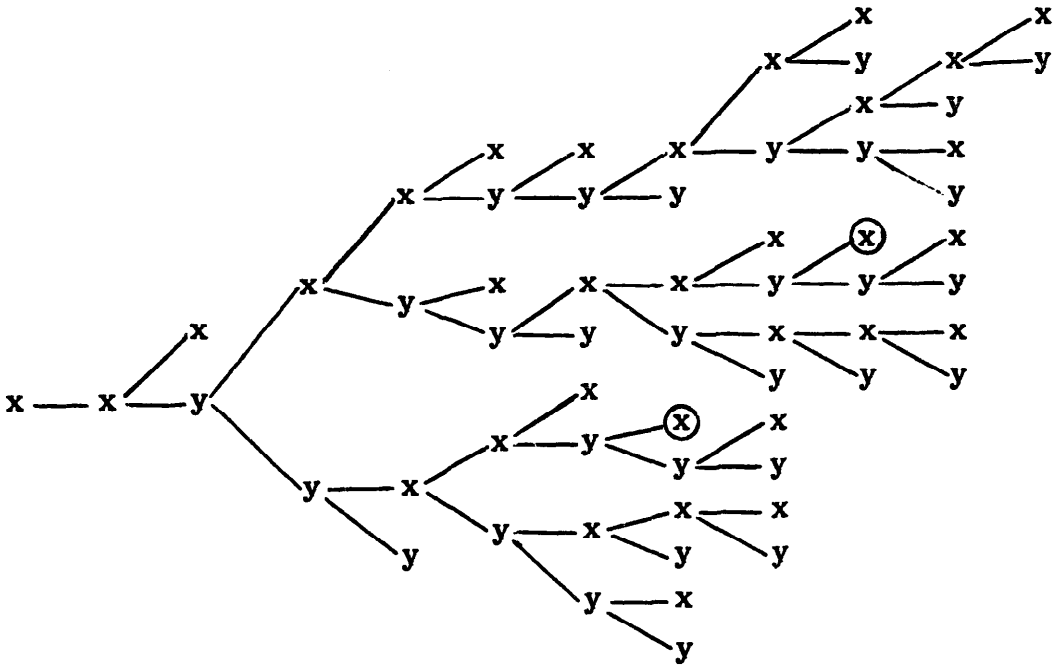


Fig. 1.

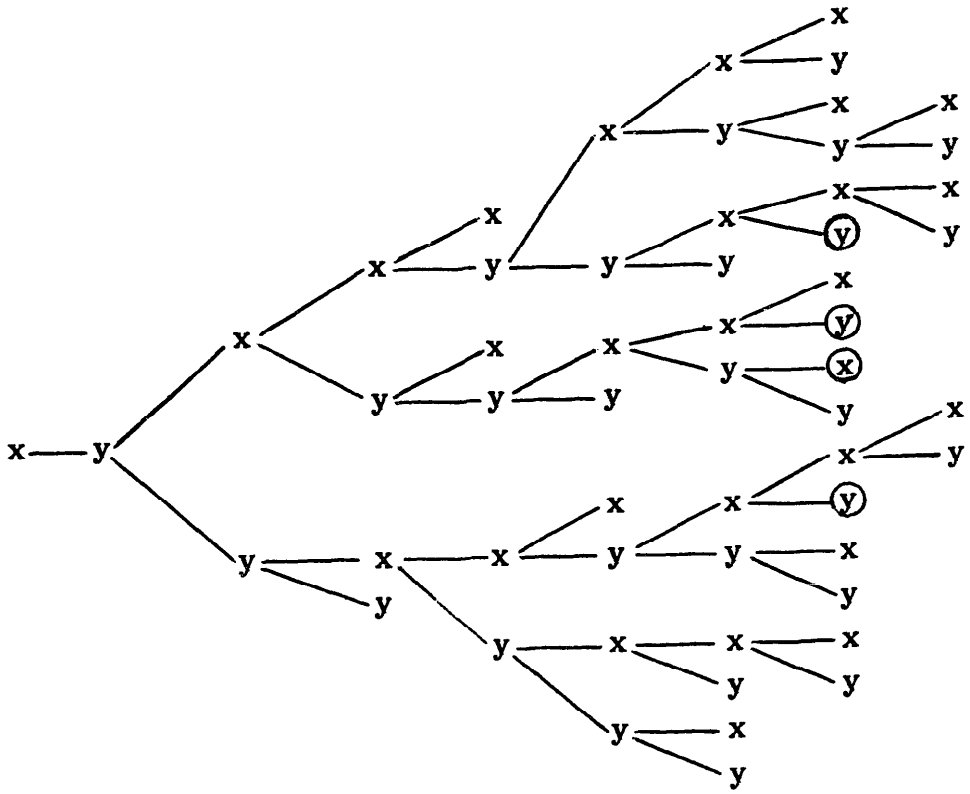


Fig. 2.

consider words up to permutations and reversal, it is sufficient to extend the words of length k by adding the letters x and y at the right.

We will construct a tree, where each path of length k beginning at the root and leading to a node represents a word of length $k+1$ consisting of the letters that are the values of the nodes of this path.

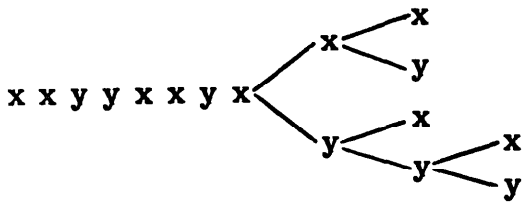


Fig. 3.

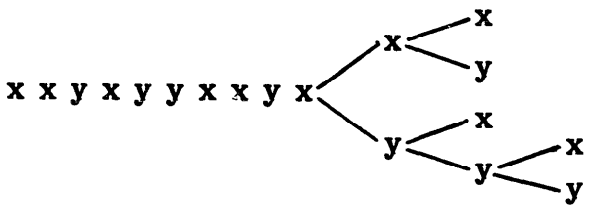
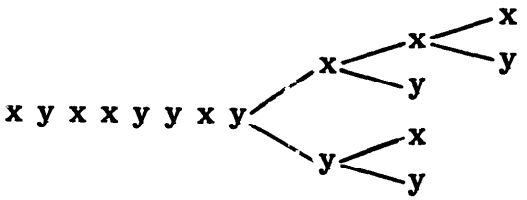
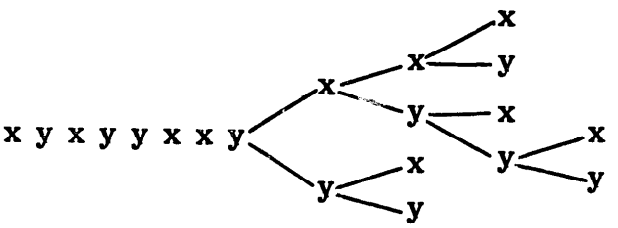
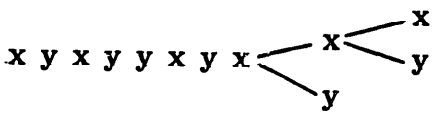


Fig. 4.







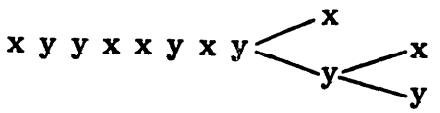


Fig. 5.

We start with the trees of Fig. 1 and Fig. 2. Figure 1 contains the words beginning with xx, whereas the words of Fig. 2 begin with xy. In Fig. 1, it remains a word of length 8, $w_1 = xxyyxxyx$, and a word of length 10, $w_2 = xxyxyyxxyx$ (cf. the circled nodes in Fig. 1); they are continued in Fig. 3 and 4.

In Fig. 2, there remain four words of length 8 (cf. the circled nodes); they are treated one by one in the same manner (see Fig. 5).

Hence, the depth of the tree beginning with xx is 12; so the maximal length of the represented words is 13. This means that the longest words beginning with xx and avoiding all patterns of X are of length 12.

The depth of the tree with root xy is 11; hence, the maximal length of words represented by the complete tree is 12.

It follows that the maximal length of the represented words is 13. □

5. Short avoidable patterns

In the final section, we will give some specific results concerning $s(p)$, where $|p| \leq 6$ and $p \in E_2^* = \{x, y\}^*$ or $p \in E_3^* = \{x, y, z\}^*$. Let $A = \{a, b\}$. We first consider $p \in E_2^*$; hence, $s(p) \leq 3$ (cf. Proposition 2.6); we consider the patterns up to permutation. Let

$$\mathrm{Av}(i) = \{p \in E^* \mid |p| = i,\ p \text{ avoidable},\ p = xp'\};$$

i.e., $\mathrm{Av}(i)$ is the set of avoidable patterns on E_2 of length i beginning with letter x.

- $\mathrm{Av}(3) = \{x^3, x^2y, xy^2\}$. Since each word on A of length 5 is divisible by x^2y, so $s(x^2y) = s(xy^2) = 3$.
- $\mathrm{Av}(4) = \{x^4, x^3y, xy^3, x^2yx, x^2y^2, xyx^2, xy^2x, (xy)^2\}$. For each pattern p containing a cube, $s(p) = 2$ holds. For the other patterns, $s(p) = 3$ since all of them divide all words on A of length 17.
- $\mathrm{Av}(5) = \{x^2yx^2, x^2yxy, xyxy^2, x^2y^2x, xy^2x^2, xyx^2y, xy^2xy, \ldots\}$
  - (i) It can be shown that $s(p) = 2$ for $p \in \{x^2y^2x, xy^2x^2, xyx^2y, xy^2xy\}$ (cf. [11]).
  - (ii) All words on A of length 39 are divisible by $p \in \{x^2yxy, xyxy^2\}$, so $s(p) = 3$.
  - (iii) There are 2^5 words on A of length 5; hence each word w on A of length $(2^5+1)\times 5$ contains two subwords of length 5 which are equal and hence contain the same square of length at most 4. These two occurrences of u are separated by at least one letter. It follows that w is divisible by x^2yx^2. This proves Proposition 3.14.
  - (iv) The patterns p on E_2 of length 5 not mentioned until now are either permutations and/or reversals of the other patterns (then $s(p)$ has been discussed) or they contain a cube or an overlapping factor. For this last case, $s(p) = 2$ holds.
- $|\mathrm{Av}(6)| = 32$; $\mathrm{Av}(6)$ contains ten patterns which are overlap-free and cube-free. There exist words on A of length 150 avoiding them; so we suppose $s(p) = 2$ for all $p \in E^+$ of length 6 (cf. [11]).

In the last part of this section, we use alph(w) to design the set of different letters of a word w: Let E be any alphabet, $w \in E^+$. Then alph(w) = $\{x \in E \mid |w|_x > 0\}$.

Now consider $E_3 = \{x, y, z\}$. The results are not so precise as in the case of E_2 since there exist much more structures on E_3. We will consider only patterns p with alph(p) = E_3.

- If $|p| = 3$, there is no avoidable pattern p with alph(p) = E_3.
- If $|p| = 4$, then no avoidable pattern p with alph(p) = E_3 is 2-avoidable, hence $s(p) \geq 3$. By writing these patterns explicitly, we see that either $s(p) = 3$ or $s(p) = \infty$ holds for them.
- If $|p| = 5$ and alph(p) = E_3, we may distinguish three cases:
  (i) The pattern p contains a cube; then $s(p) = 2$.
  (ii) The pattern p is cube-free, but contains a square. Then p contains an avoidable subword u with $|\text{alph}(u)| = 2$, and u is not 2-avoidable; the third letter serves as "boundary". Hence, $s(p) = 3$. (*Examples*: *xxyxz*, *xyzzx*).
  (iii) If p is square-free, then p is unavoidable, hence $s(p) = \infty$.
- If $|p| = 6$ and alph(p) = E_3, then there are patterns where we could not find arguments to draw conclusions about $s(p)$ as before. (*Examples*: *xyzyxz*, *xyzyyx*). Concerning $p = xyzyxz$, it is called a *commutative square*, i.e., a word of the form uv, where u is a permutation of v. All words on a 3-letter alphabet of length 8 contain a commutative square, so p is not 3-avoidable. One knows from [9] that there exist infinite words on a 5-letter alphabet avoiding the commutative square, hence p is 5-avoidable. The case of the 4-letter alphabet is an open question. So $s(xyzyxz)$ is still not determined.

References

[1] K.A. Baker, G.F. McNulty and W. Taylor, Growth problems for avoidable words, Private communication, March 1987.

[2] D.R. Bean, A. Ehrenfeucht and G.F. McNulty, Avoidable patterns in strings of symbols, *Pacific J. Math.* **85** (1979) 261-294.

[3] J. Berstel, Sur les mots sans carré définis par un morphisme, in: H.A. Maurer, ed., *Proc. of the Internat. Coll. on Automata, Languages and Programming*, Lecture Notes in Computer Science **71** (Springer, Berlin, 1979) 16-25.

[4] J. Berstel, Mots de Fibonacci, Séminaire d'Informatique Théorique 1980-81, Rapport LITP, Paris, 1981.

[5] J. Berstel, Some recent results on square-free words, in: M. Fontet and K. Mehlhorn, eds., *Proc. STACS 84*, Lecture Notes in Computer Science **166** (Springer, Berlin, 1984) 14-25.

[6] M. Crochemore, Régularités évitables, Thèse d'Etat, Rapport LITP 83-43, Paris, 1983.

[7] M. Lothaire, *Combinatorics on Words* (Addison-Wesley, Reading, MA, 1984).

[8] M.G. Main and R.J. Lorentz, An O($n \log n$) algorithm for finding all repetitions in a string, *J. Algorithms* **5** (1984) 422-432.

[9] P.A.B. Pleasant, Non-repetitive sequences, *Proc. Cambridge Philos. Soc.* **68** (1970) 267-274.

[10] A. Restivo and S. Salemi, On weakly square-free words, *Bull. EATCS* **21** (1983) 49-56.

[11] U. Schmidt, Motifs inévitables dans les mots, Thèse Université Pierre et Marie Curie (Paris 6), Rapport LITP 86-63, Paris, 1986.

[12] P. Séebold, Propriétés combinatoires des mots infinis engendrés par certains morphismes, Thèse de 3ième cycle, Rapport LITP 85-16, Paris, 1985.

[13] A. Thue, Über unendliche Zeichenreihen, *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania* 7 (1906) 1-22.

[14] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania* 1 (1912) 1-67.

[15] A.I. Zimin, Blocking sets of terms, *Matem. Sbornik* 119(161) (1982); English translation *Math. USSR Sbornik* 47 (1984) 353-364.